Supplementary Material of Subgraph-aware Graph Kernel Neural Network for Link Prediction in Biological Networks

Menglu Li, Zhiwei Wang, Luotao Liu, Xuan Liu, Wen Zhang

In this supplementary material, we first describe the definition of biological network, and introduce the details of five datasets. Further, we discuss the effect of different node attribute schemes of SubKNet, the sensitivity of several hyper-parameters of SubKNet, including the node number of graph filters (N), the maximum random walk length (P), the number of GCN layers (L), the coefficients λ_1, λ_2 and λ_3, and the number of graph filters (d_k). We also analyze the effect of subgraph size. The optimal parameter values are determined sequentially, aligning with the order of their analysis. When analyzing the influence of a certain parameter, the remaining parameters are fixed to default values (the default values N = Mean, $P = 2$, $L = 1$, $\lambda_1 = 0.5$, $\lambda_2 = 0.5$, $\lambda_3 = 0.5$, and $d_k = 32$). If the parameter has been analyzed, the default value is equal to its optimal value. The hyper-parameters we have employed are summarized in Table S1. Additionally, we provide the AUPR scores of SubKNet and baselines on cross-validation sets (Table S4) and independent test sets (Table S5), the t-SNE visualization of task-dependent and task-independent methods on the ZhangDDA dataset (Fig. S3), and the AUPR scores of SubKNet and its variants in ablation study (Fig. S4).

TABLE S1
THE OPTIMAL HYPER-PARAMETERS FOR SUBKNET

Datasets	N	P	L	λ_1	λ_2	λ_3	d_k
MDA	20	3	2	0.7	0.7	0.5	32
LuoDTI	Mean	3	1	0.5	1.0	0.2	32
ZhangMDA	20	4	2	0.5	0.7	0.5	32
ZhangDDA	20	4	1	0.5	0.5	0.5	32
PPI	20	4	1	0.5	0.5	0.5	32

I. Definition of Biological Network

To facilitate representation, we use an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ to represent both heterogeneous networks (e.g. DTI network) and homogeneous networks (e.g. PPI network). If $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ represents a heterogeneous network, the node set $\mathcal{V}$ is composed of two types of nodes $\mathcal{V}_1$ and $\mathcal{V}_2$, i.e. $\mathcal{V} = \mathcal{V}_1 + \mathcal{V}_2$. The adjacency matrix $\mathbf{A}$ can be defined as follows:

$$\mathbf{A} = \begin{bmatrix} 0 & \mathbf{A}_{hete} \\ \mathbf{A}_{hete}^T & 0 \end{bmatrix} \tag{1}$$

where $\mathbf{A}_{hete} \in \mathbb{R}^{|\mathcal{V}_1| \times |\mathcal{V}_2|}$ is the adjacency matrix of the heterogeneous network, $\mathcal{E}$ is a set of edges in $\mathbf{A}$. If $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is a homogeneous network, $\mathcal{V}$ and $\mathcal{E}$ are composed of a set of nodes and edges in this network, respectively. The adjacency matrix is $\mathbf{A} = \mathbf{A}_{homo}$, where $\mathbf{A}_{homo} \in \mathbb{R}^{|\mathcal{V}| \times |\mathcal{V}|}$ is the adjacency matrix of the homogeneous network.

II. Datasets

We assess the performance of SubKNet on five datasets: (1) The MDA dataset [1] comprises 450 microbe-disease associations, involving 292 microbes and 39 diseases. On average, there are approximately 2.719 edges per node in this dataset. Additionally, it includes a microbe taxonomic similarity matrix, which is calculated based on microbe taxonomic lineage and ranks, and a disease semantic similarity matrix [2]. (2) The LuoDTI dataset [3] contains 1,920 drug-target interactions, involving 1,493 targets and 708 drugs, four drug similarity matrices derived from drug chemical structures and three drug-related networks, and three target similarity matrices derived from target sequences and two protein-related networks. Each node approximately associates 1.745 other nodes. (3) The ZhangMDA dataset [4] consists of 4,479 miRNA-disease associations, involving 314 diseases and 412 miRNAs, one disease semantic similarity matrix, and one miRNA functional similarity matrix. On average, there are approximately 12.339 edges per node in this dataset. (4) The ZhangDDA dataset [5] comprises 18,416 drug-disease associations between 269 drugs and 598 diseases, five drug similarity matrices based on drug substructures and four drug-related networks, and a disease semantic similarity matrix. The average number of edges per node is approximately 42.482. (5) The PPI dataset [6] includes 5,581 protein-protein interactions between 2,497 proteins and a sequence-based protein similarity matrix, there are approximately 4.470 edges per node in this dataset. These datasets are described in Table S2.

TABLE S2
THE STATISTICS OF FIVE DATASETS

Dataset	Type	$\lvert\mathcal{V}\rvert$	$\lvert\mathcal{E}\rvert$	Degree
MDA	Microbe-disease association	331	450	2.719
LuoDTI	Drug-target interaction	2,201	1,920	1.745
ZhangMDA	MiRNA-disease association	726	4,479	12.339
ZhangDDA	Drug-disease association	867	18,416	42.482
PPI	Protein-protein interaction	2,497	5,581	4.470

Note: Degree is the average number of edges per node, defined as $\frac{2|\mathcal{E}|}{|\mathcal{V}|}$.

TABLE S3
RESULTS OF SUBKNET WITH DIFFERENT NODE ATTRIBUTE SCHEMES

Node attribute schemes	Multiple similarities	MDA		LuoDTI		ZhangMDA		ZhangDDA		PPI	
		AUC	AUPR	AUC	AUPR	AUC	AUPR	AUC	AUPR	AUC	AUPR
random+random	×	0.872	0.892	0.907	0.919	0.917	0.920	0.821	0.820	0.970	0.978
random+labeling	×	0.849	0.883	0.907	0.911	0.908	0.916	0.821	0.820	0.907	0.938
random+similarity	×	0.930	0.930	0.904	0.913	0.916	0.919	0.821	0.821	0.963	0.975
random+similarity	✓	-	-	0.911	0.921	-	-	0.823	0.822	-	-
random+one-hot	×	0.865	0.886	0.912	0.917	0.915	0.918	0.82	0.819	0.939	0.960
labeling+random	×	0.872	0.895	0.904	0.918	0.917	0.921	0.816	0.816	0.970	0.978
labeling+labeling	×	0.849	0.883	0.911	0.919	0.907	0.914	0.815	0.815	0.902	0.933
labeling+similarity	×	0.929	0.916	0.906	0.916	0.916	0.919	0.815	0.814	0.978	0.984
labeling+similarity	✓	-	-	0.907	0.918	-	-	0.816	0.815	-	-
labeling+one-hot	×	0.864	0.885	0.911	0.919	0.916	0.918	0.814	0.814	0.962	0.969
similarity+random	×	0.862	0.893	0.908	0.917	0.918	0.921	0.823	0.822	0.957	0.971
similarity+random	✓	-	-	0.903	0.914	-	-	0.830	0.830	-	-
similarity+labeling	×	0.836	0.873	0.907	0.912	0.909	0.917	0.822	0.821	0.905	0.941
similarity+labeling	✓	-	-	0.906	0.912	-	-	0.830	0.830	-	-
similarity+similarity	×	0.935	0.928	0.903	0.912	0.917	0.921	0.821	0.820	0.964	0.977
similarity+similarity	✓	-	-	0.909	0.916	-	-	0.830	0.830	-	-
similarity+one-hot	×	0.867	0.894	0.913	0.919	0.916	0.920	0.824	0.823	0.955	0.967
similarity+one-hot	✓	-	-	0.912	0.917	-	-	0.831	0.832	-	-
one-hot+random	×	0.873	0.897	0.910	0.922	0.918	0.922	0.841	0.841	0.949	0.968
one-hot+labeling	×	0.854	0.887	0.908	0.917	0.910	0.917	0.841	0.841	0.917	0.951
one-hot+one-hot	×	0.862	0.889	0.913	0.919	0.916	0.920	0.841	0.841	0.967	0.972
one-hot+similarity	×	**0.936**	**0.931**	0.903	0.913	**0.920**	**0.923**	0.840	0.840	**0.981**	**0.985**
one-hot+similarity	✓	-	-	**0.918**	**0.927**	-	-	**0.842**	**0.842**	-	-

Note: The highest score in each column is in bold and the second-best score is underlined. × represents this dataset only includes one type of similarity and ✓ represents multiple similarities included in this dataset. "random" is random initialization, "one-hot" represents one-hot encoding, "labeling" is node labeling, and "similarity" is node similarity. "one-hot+similarity" means that the one-hot encoding and node similarity matrix are used to obtain node attributes in subgraph-aware representation learning and graph-based representation learning, respectively.

III. Effect of different node attribute schemes

We explore the effect of node attribute schemes based on random initialization, one-hot encoding, node labeling strategy, and node similarity matrix, which are commonly used in previous studies [7]–[10]. Each scheme includes two node attributes used in subgraph-aware representation learning and graph-based representation learning, respectively. Due to the LuoDTI and ZhangDDA datasets containing multiple similarities, we also utilize the average of these similarities as the node attributes. The results are summarized in Table S3. We have the following observations: (1) Random initialization-based schemes produce sufficient performance in most datasets, indicating the stability and effectiveness of our proposed SubKNet. (2) The schemes based on multiple similarities perform better than or on par with those based on single similarity, indicating the combination of multiple similarities effectively captures diverse biological information and provides more precise quantification of the relationship between nodes. (3) The scheme (one-hot+similarity) obtains the best performance on five datasets.

IV. Parameter analysis

A. Effect of Node Number of Graph Filters

We first assess the influence of the node number of graph filters, denoted as N. This parameter plays a crucial role in determining the size and complexity of the graph filters. We vary N in $\{\text{Mean}, 1, 10, 20, 30, 40, 50\}$ while maintaining other parameters fixed. Here, Mean represents the average node number of all subgraphs in the training set. It can be observed from Fig S1 that increasing the node number of graph filters leads to an enhancement in performance. However, it is important to note that excessively large node numbers for graph filters result in a decline in model performance. This phenomenon may be attributed to the fact that large-sized graph filters introduce excessive redundant information when capturing shared information among small-sized subgraphs. Hence, we adopt $N = 20$ in MDA, ZhangMDA, ZhangDDA, and PPI datasets, N = Mean in the LuoDTI dataset for analyzing the influence of other parameters.

B. Effect of Maximum Random Walk Length

We further investigate the effect of the maximum random walk length, denoted as P. We vary P within the range $\{1, 2, 3, 4\}$. The results are shown in Fig. S1. It can be observed that setting $P = 4$ achieves the best performance on most datasets. This observation suggests that increasing the maximum length of random walks has a beneficial effect. Specifically, it enables the model to explore a wider range of substructures, which helps to learn shared information among various subgraphs. Hence, we adopt $P = 3$ in MDA and LuoDTI datasets, and $P = 4$ in ZhangMDA, ZhangDDA, and PPI datasets to analyze the effect of other parameters.

C. Effect of Number of GCN Layers

Afterward, we evaluate the effect of the number of GCN layers, denoted as L, as illustrated in Fig. S1. We consider various values for L within the set $\{1, 2, 3, 4\}$. It can be observed that node embeddings already encompass sufficient information for link prediction, even with a small number of GCN layers. It is worth noting that overly large values

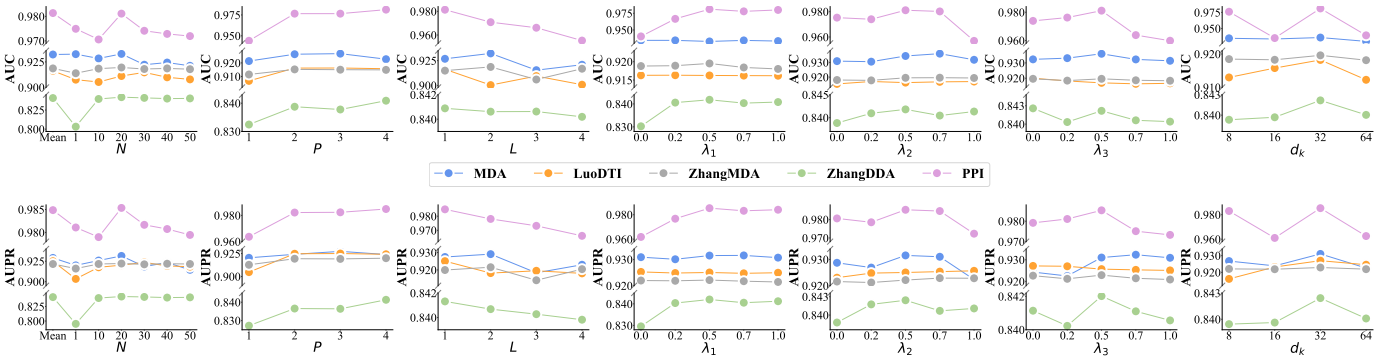

Fig. S1. Results of SubKNet with different values of the node number of graph filters (N), the maximum random walk length (P), the number of GCN layers (L), the coefficients λ_1, λ_2 and λ_3, and the number of graph filters d_k on the cross-validation sets.

for L have a detrimental effect on performance across most datasets. This phenomenon may be attributed to the over-smoothing problem inherent in GCN [11]. In this study, we adopt $L = 2$ in MDA and ZhangMDA datasets, $L = 1$ in LuoDTI, ZhangDDA, and PPI datasets.

D. Effect of coefficients λ_1, λ_2 and λ_3

We investigate the influence of coefficients λ_1, λ_2, and λ_3 by varying these coefficients in the range of $\{0.0, 0.2, 0.5, 0.7, 1.0\}$. This analysis allows us to delve deeper into the significance of diversity regularization, subgraph-aware representation learning, and graph-based representation learning. The results reveal a consistent trend on most datasets: introducing these loss terms (i.e., the coefficients are not equal to 0) tends to enhance the predictive performance of the model. These results align with that obtained in our ablation study, emphasizing the valuable contributions of all components within SubKNet to the task of biological network link prediction.

E. Effect of the number of graph filters

We assess the effect of the number of graph filters (d_k) by varying it in the range of $\{8, 16, 32, 64\}$. The results show that setting the number of graph filters to 32 achieves the best performance. The higher values of d_k produce better performance in most cases, demonstrating that various graph filters help to detect more diverse substructures for learning graph topology information. However, too large d_k may have a negative impact on performance, as lots of graph filters may introduce redundant information to fit the model.

V. Effect of Subgraph Size

To analyze the influence of subgraph size, we randomly remove a percentage of links from the training set and then implement 5-CV to assess the performance of SubKNet on the masked datasets. As shown in Fig. S2, the performance

TABLE S4
AUPR SCORES OF SUBKNET AND BASELINES ON THE CROSS-VALIDATION SETS

Categories	Methods	MDA	LuoDTI	ZhangMDA	ZhangDDA	PPI
Task-dependent methods	NinimHMDA	0.849±0.046	-	-	-	-
	MVGAEW	0.720±0.072	-	-	-	-
	IIFDTI	-	0.893±0.012	-	-	-
	GeNNius	-	0.920±0.013	-	-	-
	SFGAE	-	-	0.896±0.009	-	-
	CGHCN	-	-	0.916±0.005	-	-
	DRWBNCF	-	-	-	0.831±0.008	-
	RSML-GCN	-	-	-	0.807±0.006	-
	RAPPPID	-	-	-	-	0.883±0.014
	HNSPPI	-	-	-	-	0.956±0.003
Task-independent methods	SiGraC	0.885±0.028	0.697±0.015	0.781±0.014	0.661±0.018	0.557±0.012
	MVGCN	0.915±0.010	0.886±0.030	0.916±0.003	0.841±0.006	-
	LR-GNN	0.825±0.046	0.901±0.024	0.908±0.006	0.789±0.024	0.841±0.011
	CGCN	0.906±0.010	0.870±0.022	0.857±0.008	0.796±0.009	0.925±0.008
Subgraph-based methods	SEAL	0.897±0.010	0.923±0.010	0.921±0.003	0.825±0.006	0.883±0.010
	LGLP	0.903±0.012	0.924±0.009	0.918±0.006	0.786±0.009	0.895±0.009
	NNESF	0.878±0.017	0.843±0.016	0.747±0.010	0.679±0.003	0.894±0.008
	GCN-PS2	0.834±0.024	0.876±0.022	0.905±0.006	0.834±0.006	0.894±0.019
	SubKNet	**0.931±0.015**	**0.927±0.012**	**0.923±0.003**	**0.842±0.007**	**0.985±0.004**

Note: The bold in each column represents the highest score and the underlined denotes the second-best score. The standard deviation (±) is computed from 5-fold cross-validation results.

of SubKNet tends to decrease as more links are removed, yet the decline is less than 5% in most cases. As a percentage of links is removed, some subgraphs only contain target nodes, SubKNet hardly extracts meaningful subgraph-aware representations from these subgraphs, impacting the predictive performance of the model. However, the representations learned from the graph-based representation learning module compensate for this effect, ensuring the robustness of SubKNet.

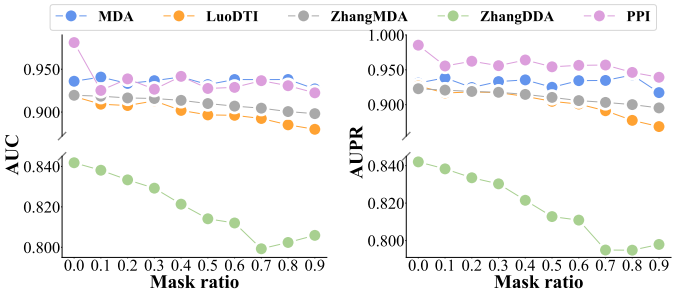

Fig. S2. Results of SubKNet with different mask ratios on the cross-validation sets.

TABLE S5
AUPR SCORES OF SUBKNET AND BASELINES ON THE INDEPENDENT TEST SETS

Categories	Methods	MDA	LuoDTI	ZhangMDA	ZhangDDA	PPI
Task-dependent methods	NinimHMDA	0.922	-	-	-	-
	MVGAEW	0.779	-	-	-	-
	IIFDTI	-	0.911	-	-	-
	GeNNius	-	0.923	-	-	-
	SFGAE	-	-	0.898	-	-
	CGHCN	-	-	0.928	-	-
	DRWBNCF	-	-	-	0.842	-
	RSML-GCN	-	-	-	0.827	-
	RAPPPID	-	-	-	-	0.941
	HNSPPI	-	-	-	-	**0.979**
Task-independent methods	SiGraC	0.919	0.645	0.821	0.653	0.592
	MVGCN	0.945	0.908	0.927	0.844	-
	LR-GNN	0.911	0.912	0.923	0.788	0.942
	CGCN	0.922	0.905	0.885	0.788	0.956
Subgraph-based methods	SEAL	0.918	0.933	0.913	0.820	0.908
	LGLP	0.943	**0.940**	0.919	0.741	0.933
	NNESF	0.914	0.852	0.734	0.669	0.932
	GCN-PS2	0.789	0.917	0.911	0.830	0.929
	SubKNet	**0.951**	0.919	**0.930**	**0.846**	0.969

Note: The bold in each column represents the highest score and the underlined denotes the second-best score.

REFERENCES

[1] W. Ma, L. Zhang, P. Zeng, C. Huang, J. Li, B. Geng, J. Yang, W. Kong, X. Zhou, and Q. Cui, "An analysis of human microbe–disease associations," *Brief. Bioinform.*, vol. 18, no. 1, pp. 85–97, 2017.

[2] Y. Ma and H. Jiang, "Ninimhmda: neural integration of neighborhood information on a multiplex heterogeneous network for multiple types of human microbe–disease association," *Bioinformatics*, vol. 36, no. 24, pp. 5665–5671, 2021.

[3] Y. Luo, X. Zhao, J. Zhou, J. Yang, Y. Zhang, W. Kuang, J. Peng, L. Chen, and J. Zeng, "A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information," *Nat. Commun.*, vol. 8, no. 1, p. 573, 2017.

[4] W. Zhang, Z. Li, W. Guo, W. Yang, and F. Huang, "A fast linear neighborhood similarity-based network link inference method to predict microrna-disease associations," *IEEE/ACM Trans. Comput. Biol. Bioinform.*, vol. 18, no. 2, pp. 405–415, 2019.

[5] W. Zhang, X. Yue, W. Lin, W. Wu, R. Liu, F. Huang, and F. Liu, "Predicting drug-disease associations by using similarity constrained matrix factorization," *BMC Bioinformatics*, vol. 19, pp. 1–12, 2018.

[6] M. Chen, C. J.-T. Ju, G. Zhou, X. Chen, T. Zhang, K.-W. Chang, C. Zaniolo, and W. Wang, "Multifaceted protein–protein interaction prediction based on siamese residual rcnn," *Bioinformatics*, vol. 35, no. 14, pp. i305–i314, 2019.

[7] C. Shen, P. Ding, J. Wee, J. Bi, J. Luo, and K. Xia, "Curvature-enhanced graph convolutional network for biomolecular interaction prediction," *Comput. Struct. Biotechnol. J.*, vol. 23, pp. 1016–1025, 2024.

[8] M. Zhang and Y. Chen, "Link prediction based on graph neural networks," in *Conference on Neural Information Processing Systems*, vol. 31, Montréal, Canada, 2018.

[9] M. Coşkun and M. Koyutürk, "Node similarity-based graph convolution for link prediction in biological networks," *Bioinformatics*, vol. 37, no. 23, pp. 4501–4508, 2021.

[10] K. Kishan, R. Li, F. Cui, and A. R. Haake, "Predicting biomedical interactions with higher-order graph convolutional networks," *IEEE/ACM Trans. Comput. Biol. Bioinform.*, vol. 19, no. 2, pp. 676–687, 2021.

[11] S. Qureshi *et al.*, "Limits of depth: Over-smoothing and over-squashing in gnns," *Big Data Mining and Analytics*, vol. 7, no. 1, pp. 205–216, 2023.

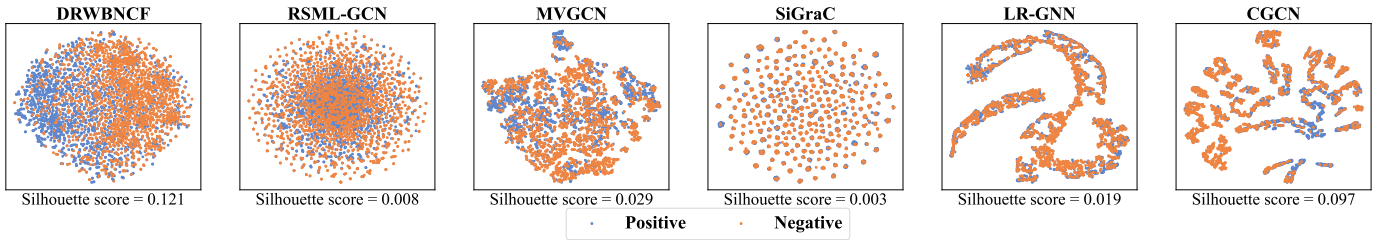

Fig. S3. The t-SNE visualization of task-dependent and task-independent methods on the ZhangDDA dataset.

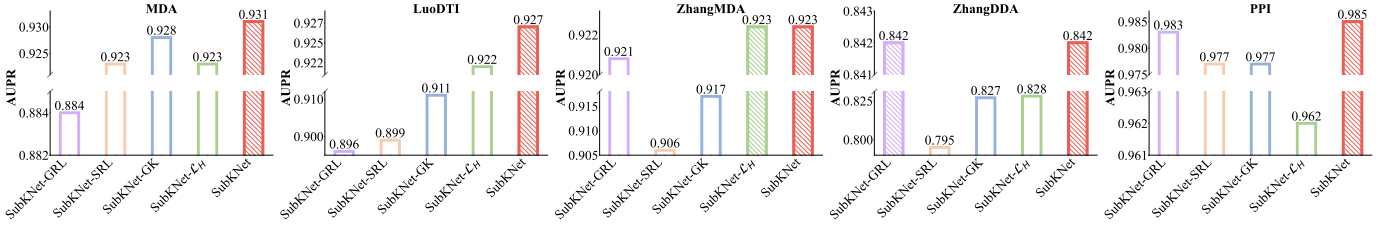

Fig. S4. AUPR scores of SubKNet and its variants in ablation study.